# Synthetic Training Data for Visual Classification Model Enhancement

Igor Overchuk

CSCI 5922, University of Colorado Boulder

## 1 Introduction

Currently, there are a variety of machine learning algorithms in use. They differ in architecture, input and output, size, and goal, but all of them need one thing to be effective. Training data. As hardware gets faster, models get larger, and scaling laws prove that models trained with more data perform better, the search for good data becomes a growing issue in this space. In my own personal experience during my undergraduate capstone project, our custom image classification model failed due to a lack of training data in a very specific domain, and we had to find another solution. In recent years, generative models have become the focus in the AI world. They offer surprisingly accurate material, whether that be text, image, or a suite of other modalities. The question this experiment aims to answer is the following: Can AI-generated images used as training data enhance the performance of visual classification models? Can we use the vast knowledge of large generative models to tune a smaller model designed for a specific task?

One solution for adding training data is data augmentation. Flipping images, cropping, or slightly altering lighting can create enough variance in a single training sample to create several valid samples. This is a great technique for dealing with scarce image data, but it has limits. There are only so many new samples that can be created before the benefits level out and the model reaches its performance ceiling. To extend beyond that, original training data is needed, and generative image AI looks like a promising source. Another option to get great results with limited training data is Few-Shot Learning with LLMs. A user can simply prompt the LLM with the training data and its labels followed by a classification question and get impressive results. This is a great concept, but an inefficient solution, as prompts to LLMs are extremely expensive to process. Finding a way to enhance a smaller model could be crucial for applications that require a lower processing demand. Notably, there is already research in the area of synthetic training data that looks very promising, but due to the recency of realistic generative models, it is still an active area of study.

The proposal is that it is possible to enhance the performance of compact visual models using synthetic image data. This is a one-time investment in expensive compute on calls to a very large model that will enable a simpler model to be put out to end users, which will reduce the compute resources used in the long run. As there is previous research on this, my experiment aims to offer a greater understanding and additional proof of concept for this idea. Generative

AI can create limitless original images in various domains. This should exceed the performance gains of data augmentation, which has a finite limit. And, as this will train a smaller fine-tuned model, it will exceed the energy efficiency of directly calling LLMs.

## 2 Related Work

*Using Generative AI to Create Synthetic Training Data to Enhance Models*

- [1] This Dataset Does Not Exist: Training Models from Generated Images
- [2] AI-Generated Images as Data Source: The Dawn of Synthetic Era
- [3] A novel algorithm can generate data to train machine learning models in conditions of extreme scarcity of real world data

This study fits into this topic. These papers explore various methods/use cases of training visual models with synthetic data. The first explores training on synthetic data alone with promising results. My study aims to test various combinations of training subsets to see which can be the most effective. The second article explores generating images for edge case models without a lot of real world data. This is similar to what is being studied here, but my focus isn't on a specific edge case, it is more general. The third explores the design of an additional genetic algorithm to improve the output of generative models in cases of extreme scarcity. In this study, breast cancer detection. My study does not aim to improve synthetic image generation, rather it looks to use existing generative models as they are.

*Model Degradation When Training on AI Generated Data*

- [4] AI models collapse when trained on recursively generated data
- [5] On the caveats of AI autophagy

These articles warn us of the inevitable mixing of synthetic and real-world data out on the internet, where most large models get their data from right now. They study the performance of these models as more and more synthetic data is injected into the training process and find that eventually the ratio of real/synthetic data becomes too great, and the model collapses. They focus mainly on LLMs, so my study focusing on image models differs, but I'll be monitoring for potential degradation as the percentage of synthetic data increases as well.

*Ethical Implications of Using Synthetic Data*

- [6] The Use of Synthetic Data to Train AI Models: Opportunities and Risks for Sustainable Development
- [7] Machine learning and the politics of synthetic data

These articles focus on the ethical and political implications of training models with synthetic data. Specifically, how biases from the GANs used to supply the generated data can propagate into the models that the data is used to train. Also, they talk about the issues of synthetic data ownership when using these GANs. My study does not dive into these ethical considerations in depth, but it is very important to consider these points as they can have significant implications.

## 3 Methodology

I will use a simple domain that has limited training data to simulate a scarce data environment with a manageable task. The dataset is from Kaggle. It has 90 different animal categories, with 60 images total per animal. That makes 5400 images in the entire dataset, small by today's standards. I will split the dataset into 30 training images and 30 test images to create a more scarce training environment, as 30 images per category is unlikely to be enough to achieve great results with small models.

To generate the synthetic images, I will be using ComfyUI, a software application that is installed locally on your machine and uses your graphics card to run a generative model of your choice to create images from prompts. It also has other features, but I will only use this one. The generative model I am using is Stable Diffusion 1.5. Using my laptop NVIDIA RTX 2060 GPU, it takes about 5 seconds to generate one 512x512 image. My goal is to generate at least 30 synthetic images per category to achieve a 50/50 real/synthetic split in my training set. Alternatively, I explored using state of the art models, such as ChatGPT's image generation and Midjourney, but their daily limits would make it difficult to create a meaningful amount of additional images.

The models I will use to test performance are a simple CNN, to see how a really small model reacts to the extra data, and a combined pre-trained model using ResNet50, to see how a larger capacity model reacts. If time permits, I would also like to test a visual transformer to see if it has a different accuracy curve, but I will start with the other 2 first. For the simple CNN, I will use a model with at least 5 convolutional/max pool layers, standard regularization techniques like L2 norm and dropout, and a fully connected layer at the end for the classification. For the pre-trained model, I will freeze the pre-trained weights of ResNet50 to speed up training, use the same regularization techniques, and output the classification with a fully connected layer at the end. Exact architectural/training details of both models will be determined during hyper parameter tuning on just the real training data. Once the best relative performance is found, all hyper parameters will be kept constant for testing with various amounts of synthetic data.

## 4 Experiments

This study will perform a single experiment to see how various amounts of synthetic training data affect performance on a test set of real images. As mentioned

above, the real data will be split into 30 training and 30 testing images. The models will be tuned using real data only, and once the optimal parameters are found, they will be kept constant for all other training runs. There will be a recorded baseline accuracy on only the real training data for each model. Then, 15, and 30 synthetic images will be inserted into the training pool for each category. This will create a 2x3 grid of results. 2 rows for each model, and 3 columns for 0% synthetic, 50% additional synthetic, and 100% additional synthetic. The metrics used to evaluate each model will be the basic accuracy, precision, recall, and F1 score. In the end, an analysis will be performed on the kind of curves each of these metrics produces with different combinations of training datasets to determine the feasibility of improving model performance with synthetic data.

## References

1. Besnier, V., Jain, H., Bursuc, A., Cord, M., Pérez, P.: This dataset does not exist: training models from generated images. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE (2020) 1–5
2. Yang, Z., Zhan, F., Liu, K., Xu, M., Lu, S.: Ai-generated images as data source: The dawn of synthetic era. arXiv preprint arXiv:2310.01830 (2023)
3. Niel, O.: A novel algorithm can generate data to train machine learning models in conditions of extreme scarcity of real world data. arXiv preprint arXiv:2305.00987 (2023)
4. Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., Gal, Y.: Ai models collapse when trained on recursively generated data. Nature **631**(8022) (2024) 755–759
5. Xing, X., Shi, F., Huang, J., Wu, Y., Nan, Y., Zhang, S., Fang, Y., Roberts, M., Schönlieb, C.B., Del Ser, J., et al.: On the caveats of ai autophagy. Nature Machine Intelligence (2025) 1–9
6. Marwala, T., Fournier-Tombs, E., Stinckwich, S.: The use of synthetic data to train ai models: Opportunities and risks for sustainable development. arXiv preprint arXiv:2309.00652 (2023)
7. Jacobsen, B.N.: Machine learning and the politics of synthetic data. Big Data & Society **10**(1) (2023) 20539517221145372